Linguistic structure evolves to match meaning structure

Linguistic Structure Evolves to Match Meaning Structure

Mónica Tamariz (monica@ling.ed.ac.uk)
Language Evolution and Computation Research Unit, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 1AD, UK

Abstract

Quantitative analysis has usually highlighted the random nature of linguistic forms (Zipf, 1949). We zoom in on three structured samples of language (numerals; playing cards; and a corpus of artificial languages from Kirby, Cornish & Smith 2008) to quantitative explore and illustrate the idea that linguistic forms are nonrandom in that their structure reflects the structure of the meanings they convey. A novel methodology returns frequency spectra showing the distribution of character *n*-gram frequencies in our language samples. These spectra, purely derived from *linguistic form,* clearly reflect the quantitative structure of the underlying *meaning spaces*, as verified with a new information theoretical metric of compositionality. Moreover, analyses of a diachronic corpus of languages show that linguistic structure *gradually adapts* to match the structure of meanings over cultural transmission.

Keywords: frequency distributions; form-meaning systematicity; cultural language evolution.

Introduction

Linguistic forms are quantitatively structured as illustrated by the facts that lexical item frequency and regularity are inversely correlated (e.g. Bybee & Hopper, 2001); that the frequency of a word is inversely proportional to its frequency rank following a power law (Zipf, 1949); or that word type-token ratios and lexical diversity are used to measure text complexity (Laufer & Nation, 1995). The structure of linguistic forms has also been shown to reflect, to small but statistically significant extents, the structure of the meanings that language conveys. This is most obvious in morphosyntactic paradigms, where words that share an affix also share a grammatical meaning such as tense, aspect, gender or number. However, lexical phonology has also been shown to reflect semantic structure in phonaesthemes (Bergen, 2004) or through sound symbolism (Sapir, 1929; Hinton, Nichols, and Ohala, 1994). Moreover, the phonology of lexical roots has also been shown to predict their syntactic categories (Monaghan, Christiansen and Chater, 2007) and, for the whole language, words that sound similar tend to have similar distributional (syntactic and semantic) properties in speech (Shillcock et al., 2001, Tamariz, 2008). The systematic relationship between forms and meanings means that, given access to the structure of *forms,* we can know something about the structure of the corresponding *meanings*. The first novel method introduced in this paper specifically seeks to discover quantitiative information about meaning spaces by looking at the frequencies of *n*-grams in linguistic forms.

The correlation between form and meaning structure is in many cases compositional in nature. In a compositional system, the meaning of a complex signal depends on the meanings of its component simplex signals and the rules used to combine them, e.g. the meaning *impenetrable* depends on the meanings of root *penetr* and affixes *im* and *able* as well as the way these are put together. Cornish, Tamariz & Kirby (2010) introduced a method to quantify the *details* of compositionality of artificial languages. The second novel method we introduced is a metric yielding a *single measure* of the compositionality of a system. This is used to quantify, from *form and meaning* information, the *compositionality* of a language.

The two above-mentioned methods are applied to two samples of natural language and one corpus of artificial languages where the highly structured meaning space is known. First, numerals 1-999 and the names of playing cards are analyzed to illustrate (a) how the distribution of *n*-gram frequencies can reveal meaning structure based on form structure in extant language and (b) the metric of compositionality. Second, a diachronic corpus of artificial miniature languages (from Kirby, Cornish & Smith, 2008) is analyzed to show the *process* of change of linguistic form structure to match meaning structure, thus directly testing the hypothesis that languages adapt to the structure of meanings over cultural transmission.

1. Spectral and Compositionality analysis of extant language samples

Methods

The frequency spectrum of a linguistic sample will reveal quantitative structure in linguistic forms. We obtain the spectra of numeral types 1-999 and playing card names to illustrate the method. These samples refer to meanings with known clear quantitative structure; additionally, in the samples, certain characters strings occur very frequently, e.g. “six” or “hundred” in the numerals and “queen” or “spades” in the card names. Knowing the meaning spaces, we expect the string “queen” to occur four times in the card name list, and the string “spades” to occur thirteen times. Indeed, frequencies four and thirteen should be very prevalent in the list of playing card names, because in a real deck of cards there are four suits and thirteen number and face cards. In contrast, in a matching list of words referring to 52 random objects we would not expect particular strings

to recur to the same extent; we would be even more surprised to find particular string *frequencies* being especially prevalent. In fact, for the random list we would expect low frequencies to be very prevalent (frequent) and high frequencies to be very rare, and this inverse relationship should follow a power law (Manning & Schütze, 1999). This prediction is tested by looking at the *n*-grams (uni-, bi- and tri-grams aggregated) in the words: For the frequencies of *n*-gram frequencies of a set of random words, the resulting spectrum should follow a power law. But for one of our special samples, the resulting spectrum should reflect the structure of the *meanings* that the lexical set refers to. A Monte Carlo analysis is used to calculate how different the spectra obtained with our language samples are from those obtained with random words.

Additionally, we have full knowledge of the meaning spaces underlying these two samples, and of the mappings between those meanings and the forms are in use (e.g. the form "ace of spades" is used to refer to the card depicting a single spade). We expect that, for these highly structured meaning spaces, the mappings between forms and meanings will be compositional in nature. Another Monte Carlo analysis tells us whether the mappings between signals and meanings are significantly compositional.

Materials

The first sample comprises English numerals for 1-999, removing any spaces between words; for instance, 541 is "*fivehundredandfortyone*". For the playing cards, similarly, the names with no spaces are also used, e.g. "*jackofspades"*.

The random language samples for the Monte-Carlo analyses contain the same number of items as the corresponding target list (numerals or cards). Each item starts with one word randomly selected from the spoken section of the British National Corpus[1]. It continues with the following word in the corpus, then the next one and so on until the item has the same number of characters as the corresponding item in the target list (no spaces here either).

Spectral analysis

For the spectral analysis, all *n*-grams were extracted from each sample and their frequencies counted. The frequencies of frequencies were then computed. First, we examine the fit to a power law by comparing the fit (R^2) and slope (b) of the power law regressions of the target versus the random language samples. Regressions are calculated on the set of *n*-gram frequencies (x) and their frequencies (y). We expect significantly lower R^2 and higher b values for the target samples, indicating that their frequency structure is different from those in random linguistic items. Second, we construct a spectrum based on the *n*-gram frequency structure of the sample. For each *n*-gram frequency, we obtain ands plot its *z*-score by comparing its frequency in the sample against 1,000 random samples; (z-scores are used throughout the paper since all random distributions in the Monte Carlo analyses were approximately normal). Spectra thus show, for each n-gram frequency, how divergent it is from what would be expected in random linguistic sample. If our hypothesis is correct, these *z*-scores should match aspects of the quantitative structure of the meaning space expressed by the forms in the sample.

Compositionality analysis

For the compositionality analysis, *RegMap* (Tamariz & Smith, 2008; Cornish, Tamariz & Kirby, 2010; Tamariz, 2011) was used. This metric of the *Reg*ularity of the *Map*pings involves, crucially, *segmenting* the meanings and signals. Meanings are segmented into simplex meaning features (for the numerals, hundreds, tens, units; for the playing cards, suit and number). Signals are divided into meaningful segments (numerals are divided into three segments, one each for units, tens and hundreds, so for "twentyseven" we have *Ø*, *twenty* and *seven*; playing card names are divided into two segments, just before "of", so for "queenofhearts" we have *queen* and *ofhearts*). Then, we obtain *RegMap* for each meaning feature - signal segment pair.

$$\text{Re}gMap = \sqrt{\left(\frac{1-H(s\mid m)}{\log(n_s)}\right)\times\left(\frac{1-H(m\mid s)}{\log(n_m)}\right)} \quad (1)$$

RegMap (Eq. 1) is based on information theory conditional entropy H(A|B), which yields the amount of uncertainty, or surprise, that two features are associated; in this case, for instance that a form segment s (e.g. the first segment in the numeral) is associated with a meaning feature m (e.g. the units), after having seen all the system (e.g. after having learned the name of all playing cards). The conditional entropy of signals given meanings and of meanings given signals are both taken into account, since they are not symmetrical; they are normalized and subtracted from 1 to return levels of confidence or reliability of the association, rather than of uncertainty.

For a language with N meaning features and M signal segments, we obtain an N x M matrix of *RegMap* values. Fig. 1 illustrates this for the numerals. High values indicate that variants of the segment reliably predict the variants of the meaning feature. So, for the pair (Segment 1, hundreds) we obtain the highest value, since the first segment {nil, onehundred, twohundred, …, ninehundred} perfectly predicts the hundreds {0, 1, 2,… 9}. For (Segment 3, units) *RegMap* is somewhat lower, reflecting the presence of exceptions – 11 to 19 are irregular in this respect, the last segment of the numerals does not express the units. Low values indicate low predictability.

	hundr	*tens*	*units*
Segm1	1.000	0.018	0.017
Segm2	0.000	0.959	0.175
Segm3	0.127	0.000	0.910

Figure 1. Matrix of *RegMap* values for the three signal segments and the three meaning features in the numerals 1-999. As expected, the first segment reliably predicts the hundreds, the second the tens and the third the units. While *RegMap* is perfect for the hundreds, the values for tens and units indicate the presence of exceptions there.

1 Data extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts used are reserved.

Compositionality is calculated by applying the same algorithm to the matrix of *RegMaps* obtained for all combinations of meaning features and signal segments (e.g. for the numerals, to the matrix shown in Fig. 1). In a highly compositional system, each segment is reliably associated (high *RegMap*) with *one and only one* meaning feature, and badly with the others, and this is reflected in *Comp* (Equation. 2).

$$Comp = \sqrt{\left(\frac{1-H(S \mid M)}{\log(n_S)}\right) \times \left(\frac{1-H(M \mid S)}{\log(n_M)}\right)} \quad (2)$$

Here *S* refers to signals and *M* to meanings in the language; *Comp* measures the reliability of the one-to-one association between the signal segments and the meaning features in the language overall. The significance *Comp* values is assessed with a Monte Carlo analysis.

Results

Table 1. Results of the Monte Carlo analysis, showing the fit (R^2) and beta coefficient (*b*) of a power law regression for the *n*-gram frequency distributions in the numerals and playing card names.

	R^2		*b*	
	Num	*Cards*	*Num*	*Cards*
Value	0.208	0.372	-0.290	-0.773
Mean (N=1,000)	0.722	0.801	-0.971	-1.365
S.D. (N=1,000)	0.020	0.032	0.022	0.032
z-score	-26.395	-13.320	30.394	18.600
p value	.000	.000	.000	.000

The frequency-of-frequency distributions both in the random samples and in our structured samples were best explained by power law regressions than by linear, logarithmic, polynomial or exponential regressions. Table 1 shows, however that the distributions in our target samples are significantly worse fitted by power law regressions than the random samples and their regressions have also significantly different *b* values, indicating that the structured samples have flatter regression curves, with less frequent low frequencies (e.g. no *n*-grams occur only once in the card name list) and more frequent high frequencies (e.g. the frequencies of the *n*-grams in "spades" in the cards) than in the random samples.

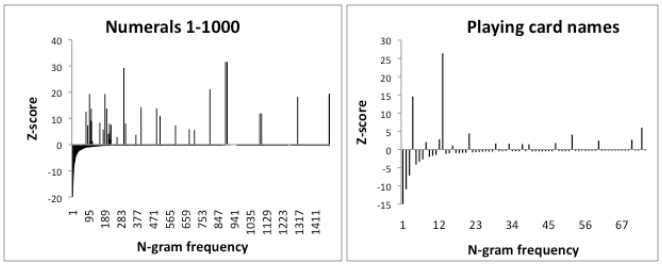

Figure 2. Spectra of the numerals and playing card name samples: *Z*-scores[2] of the *n*-gram frequencies.

[2] Absolute *z*-score values greater than 1.96 correspond to a 0.95 confidence level and greater than 3.29, to a 0.999 confidence level.

The spectra in Fig. 2 shows that, in the playing card list, *n*-gram frequency values 13, 4, 73, 21, 52, 12, 70, 60 and 8 return significantly positive *z*-scores. These values are clearly related to the underlying meaning space. Inspection of the *n*-grams with frequency 13, for instance, illustrate their significance in the meaning set of playing cards: (*ofs, fs, fsp, sp, p, spa, pa, pad, ad, ade, d, de, des, es, es*); (*ofc, fc, fcl, cl, c, clu, lu, lub, ub, b, bs, bs*); (*ofh, fh, fhe, he, hea, ea, ear, ar, art, rts, rt, ts, ts*); (*ofd, fd, fdi, di, dia, ia, iam, am, amo, m, mo, mon, on, ond, nd, nds, ds, ds*). The spectrum of the numerals is analyzed in Table 2.

Table 2. *N*-gram frequencies with significant positive *z*-scores in the numerals.

Freq	*z*	*Freq*	*z*	*Freq*	*z*	*Freq*	*z*
891	31.62	400	14.36	108	9.16	680	5.89
900	31.62	490	13.79	112	8.87	180	5.79
300	29.27	200	13.75	160	8.33	710	5.59
800	21.07	110	13.71	310	8.03	224	5.36
100	19.33	80	12.57	216	8.00	210	4.14
190	19.28	1090	11.91	225	7.61	370	3.82
1500	19.06	1100	11.91	600	7.39	260	2.89
1310	18.23	510	11.00	90	7.35	220	1.99

A first glance at Table 2 shows the abundance of multiples of 10, indicating a reflection of the decimal system. However, a closer inspection reveals subtleties relating to the precise structure of the sample, including the fact that it goes up to three levels (units, tens and hundreds). At the top of the rank we find *n*-gram frequencies 900, 891 and 300. A closer look at the precise *n*-grams that have these frequencies illustrate their significance. Nine *n*-grams have frequency 900 (*hu, hun, un, und, ndr, dr, dre, red, ed*); six *n*-grams have frequency 891 (*eda, da, dan, a, an, and*); and 22 *n*-grams have frequency 300 (*tw, w; fo; fi; so, six, ix, x; se, sev; ei, eig, g, igh, ig, gh, ght, th; ni, nin, in, ine*). This tells, us, for example, that exactly one word, "hundred" occurs precisely 900 times in the numeral sample; the sequence "edand", a subset of "hundred and" occurs 891 times; and the unique digit roots for 2, 4, 5, 6, 7, 8 and 9 occur 300 times each (100 times as units plus 100 times as tens plus 100 times as hundreds).

Table3. Compositionality values for the numerals and playing card names and their significance values.

	Num	*Cards*
Comp	0.672	1.000
Mean (*N*=1,000)	0.035	0.154
S.D. (*N*=1,000)	0.031	0.049
z-score	20.581	17.271
p value	0.000	0.000

Table 3 shows the results of the *RegMap-Compositionality* study. As expected, these two samples return much higher compositionality levels than chance would predict.

Discussion

These results show how the structure of meanings in highly organized, closed semantic sets can be detected in the quantitative structure of the linguistic items that refer to them. Significant departures from a power law distribution of the frequencies of character *n*-gram frequencies indicate structure in the samples, and this is confirmed by their highly significant compositionality values. Finally, inspection of the *n*-grams with high-frequency frequencies in the spectra confirms that the structure found in the linguistic form samples corresponds to structural features of the meaning space.

The frequency analyses of the two language samples share three features. First, the most salient frequencies in the spectra give us an idea of the quantitative structure of the underlying meaning space. Second, we find few low frequency *n*-grams, in fact a lot fewer than expected by chance in random samples. This indicates that existing *n*-grams tend to be reused. A structured meaning space, by definition, is organized along features (such as number, suit, but also tense, case etc) that are shared by several items. Correspondingly, the forms associated to such a meaning space contain many repetitions of the *n*-grams expressing the common features. Third, the language samples tend to be efficiently structured. We find little ambiguity, with many of the *n*-grams corresponding to meaning features being unique to them, suggesting that the systems are adapted to allow maximal distinction between variants of the same feature (e.g. numerals for 0-9 are maximally distinct). On the other hand we find *n*-grams occurring exactly once in every item in the list, such as "of" or final *s* in the card names. These may help identify members of the meaning space: the template "*x* of *x*s" in the appropriate context signals the name of a card – any card.

Our samples are admittedly extreme cases unequivocally quantitatively structured meaning sets. Nevertheless, these results suggest an avenue to explore form-meaning correspondence quantitatively. The methods can arguably be adapted, refined and extended to detect subtler correlations in larger, less organized language samples.

We now turn to the question of how this correspondence could have come about.

2. The evolution of meaning-form compositionality

The previous studies provided evidence for a measurable match between linguistic form and meaning structure. Such nonrandom, efficient and economical correlations are likely to be the product of either intentional design or a selection process. We cannot rule out intentional design in the two analyzed samples. We can, however, investigate whether a process of selection and adaptation could result, over time, in such well matched form-meaning systems.

Materials

The novel methods described above were applied to data collected by Kirby, Cornish and Smith (2008) (henceforth, KCS). They carried out an artificial language learning study involving a highly structured meaning space. In the experiments reported in that paper, participants had to learn artificial languages used to name 27 objects, which combined three shapes, three colours and three motions. The initial names for those objects were randomly constructed out of CV syllables, and consequently there was no strong match between the structure of forms (names) and the structure of meanings (objects). One participant was trained with 14 items out of this "random" system and then tested in the following way: when presented with each of the 27 objects they had to type the name they thought corresponded to it. Importantly, each participant would be trained on half of the items produced by the previous one. The languages change and, after ten such iterations, the names are no longer random but their structure reflects the structure of the meanings. They collected in this way eight language chains which constitute a perfect corpus to track the process of adaptation of linguistic forms to the structure of the meanings. The output languages produced by each of the participants (at each "generation") are analyzed.

KCS reported two experiments, the second of which introduced an extra manipulation. The selection of the 14 items of a language to go in the next participant's training set was not random, but explicitly excluded homonyms, that is, items that had been given the same name. The four language chains in the first experiment evolved to display "structured underspecification", with high degrees of homonymy (in the extreme, a couple of language chains ended up with only two words to name all 27 objects). The four language chains in the second experiment, having undergone the "homonymy filter", evolved to display compositionality. Our *n*-gram analyses were applied to all eight languages chains; the *Comp* analyses, for reasons explained in the following section, were only carried out on the four language chains in the filtered condition.

Methods

We performed a spectral analysis (see page 2 above) on all languages in KCS's studies. The fit to a power law regression is expected to decrease over generations, reflecting a progressive departure from randomness. Given the structure of the meaning space, where each feature (each of the three colours, motions and shapes) is present in nine objects, *n*-gram frequency 9 is predicted to be the most salient in the spectrum for the final, more adapted languages. For the Monte Carlo analysis, we compare the *n*-gram frequencies in the language at each generation with those in 5,000 random languages, generated in the same way as KCS created their initial, random languages.

The *RegMap-Comp* analysis is carried out only for the languages in the filtered condition of KCS's studies to quantitatively reveal the process of gradual adaptation of the language structure to the meaning space structure. The 27

words in each language are segmented into three meaningful chunks following the methods set up in in Cornish, Tamariz and Kirby (2010); three meaning space dimensions (colour, shape and motion) are considered. *RegMap* analyses are run to measure the regularity of the mappings between each segment and each meaning dimension at each generation. *Comp* is then calculated at each language-generation to reveal the evolution of compositionality. The four language chains in KCS's unfiltered condition were not used for these analyses because words were not amenable to any meaningful segmentation. The significance of *Comp* is assessed, as before, with a Monte Carlo analysis involving 1,000 randomisations of the target language. Random languages were constructed by scrambling the mappings between the signals and meanings.

Results

The results in Fig. 3 (left) indicate that the frequency of frequency distributions in the initial, random languages have good fits to power law regressions, with R^2 values close to 1 (indicating that they are indeed random). As expected, these values decrease as the languages are learned and reproduced by successive participants (generations), suggesting that they become more structured. In Fig. 3 (right) it is apparent that the slopes of these regressions tend to flatten out in the later generations, indicating as before that there are less *n*-grams with lower frequencies and/or more with higher frequencies than in the early languages. Paired *t*-tests return significant differences between the R^2 and *b* values in the initial and final generations (for R2, t=7.54, p=0.000; for the slopes t=7.30, p=0.000).

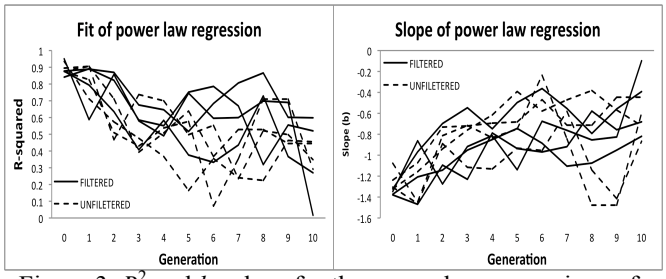

Figure 3. R^2 and *b* values for the power law regressions of the *n*-gram frequency of frequency distributions in the eight languages from KCS.

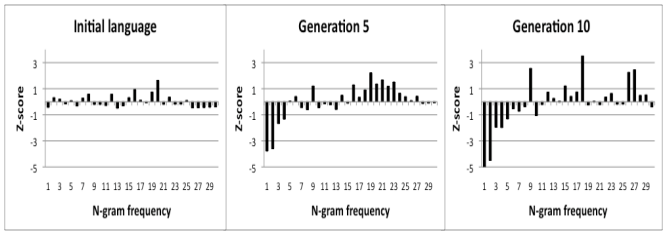

Figure 4. Three spectra illustrate evolution of form structure over time: Average *Z*-scores of the *n*-gram frequencies from all eight languages in Cornish, Kirby and Smith (2008) at generations 0 (initial languages), 5 and 10.

Fig. 4 shows how the spectra based on *n*-gram frequency distributions in KCS's languages change over the generations. Initial spectra show no significant departures from chance (no *z*-score has an absolute value greater than 1.96). At later generations, lower *n*-gram frequencies become significantly lower than expected by chance, while a few higher frequencies (namely 18, 9 27 and 26) have significantly positive *z*-scores. This result confirms the expectation that frequency 9 would be the most salient for these *forms* because each *meaning* feature appears in 9 items in the language. It also indicates high re-use of units and, more importantly, a *gradual process of adaptation* of the language from randomness towards a good match of the meaning space structure.

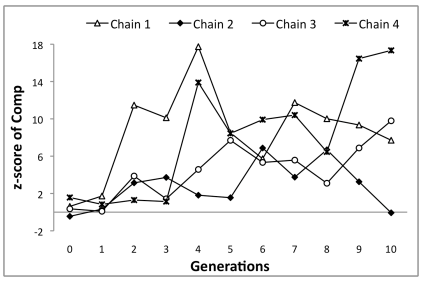

Figure 6. *Z*-scores of *Comp* values at each generation of the four language chains from Kirby, Cornish and Smith (2008), (filtered condition).

Fig. 6 shows that *Comp* tends to increase over time to reach significantly high levels. Initially random, the mappings between features of form and features of meaning become more one-to-one as the languages are repeatedly learned and produced. This strongly suggests that the linguistic form structure in the later generations revealed in the fit to power-law regressions and the spectra is actually related to meaning structure.

Discussion

The spectral analysis of the KCS data reveals how initial, randomly constructed lexical items gradually acquire a quantitative structure that matches the structure of the meanings that those lexical items denoted. This happens progresssively, as the language is repeatedly transmitted to new participants. By generation 10 the spectra share the three features observed in the numeral and playing-card names spectra. First, the relationship between the most salient frequencies and the meaning space: KCS's meaning space is comparable to playing cards in the sense that it comprises all possible items given the three colours, shapes and motions. The most significant frequencies, 18, 9, 27 and 26, reflect on the one hand the fact that there were nine items of each colour, shape and motion and that sometimes only one of those values was expressed in the language, with e.g. the 9 red objects denoted by a name starting with "po" and all other 18 denoted by a name strating with "wa". On the other hand, frequencies 27 and 26 indicate that

(nearly) all 27 names in a language shared some *n*-grams. For instance, in language chain 1, which attained a high degree of compositionality, the penultimate character was "k" in all words. This character could be said to have taken on the function of identifying membership of the language.

Second, the final languages have significantly fewer low-frequency *n*-grams than expected by chance, again indicating repetition of a small number of *n*-gram types. Third, efficient structure: repeated *n*-grams are not randomly distributed. At generation 10, languages tend to have a unique *n*-gram devoted to each meaning feature, and these *n*-grams are re-used and recombined according to the features of the object to be named.

Discussion and conclusions

Frequency analyses of large linguistic corpora have stressed the random, unpredictable nature of language structure, as reflected in power-law distributions (Zipf, 1949). By zooming in on small language samples whose associated meanings are very structured, we asked: Does the frequency distribution of sublexical units in a word sample reflect quantitative properties of the meaning space associated with those words? In our selected samples, as expected, this seemed to be the case. Discovering quantitative regularities in linguistic forms may therefore indicate that the corresponding meanings are quantitatively structured. Conversely, we can predict that when a quantitatively structured meaning space is expressed linguistically, traces of that quantitative structure should be detectable in the linguistic forms.

Adding an evolutionary dimension, we asked: How did linguistic form-meaning mappings become compositional? Our analyses of diachronic samples of artificial language chains suggest that the strong correlation between form and meaning structure is, at least in part, the result of a *process of adaptation* of forms to the structure of the meaning space.

This highlights meanings as a causal factor in linguistic structure and emphasizes the interplay between meaning and form structure during language learning and evolution. The information-theoretical basis of the *RegMap* and *Compositionality* metrics indicates the important role of learning principles such as efficiency and economy in the adaptation process. The resulting languages tend to be optimally compressible: they contain the minimum number of distinct meaningful units and recombination rules required to express all the meanings.

The evidence presented also highlights the fact that inference of linguistic structure by learners is driven by regularities in their input. Structure in the forms, such as repetition of the same *n*-gram in all words and a nonrandom *n*-gram spectrum, and structure in the form-meaning mappings, such as consistent cooccurrence between *n*-grams and meaning dimensions, seem to be especially salient to learners. Regularities are then not only well remembered and employed to name learned items, but also generalized to name novel items.

Finally, one word on the methodology. Spectral analyses capture and can help visualize frequency structure in linguistic forms not just with character *n*-grams, but at any level. *RegMap* and *Compositionality* metrics are also able to capture meaning-form regularity at any degree of analysis, by defining the form segments and meaning features relevant to our research questions.

Acknowledgments

This work was supported by UK AHRC Grant AH/F017677/1.

References

Bergen, B. (2004). The psychological reality of phonasethemes. *Language*, 80(2).

Bybee, J. and Hopper, P. (2001). *Frequency and the Emergence of Language Structure.* Amsterdam: John Benjamins.

Cornish, H. Tamariz, M. and Kirby, S. (2010). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. Special issue on Language as a complex Adaptive System. *Language Learning: 59:4S1*.

Hinton, L., Nichols, J. & Ohala, J. J. (1994). *Sound symbolism*. Cambridge: Cambridge University Press.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative Cultural Evolution in the Laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105(31).

Laufer, B. and P. Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16.

Manning, C. & Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Monaghan, P., Christiansen, M.H., & Chater, N. (2007). The Phonological Distributional coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, 55.

Sapir, E. 1929. A Study in Phonetic Symbolism. Journal of Experimental Psychology, 12.

Shillcock, R.C., Kirby, McDonald, S. & Brew, C. (2001). Filled pauses and their status in the mental lexicon. Proceedings of the 2001 Conference of Disfluency in Spontaneous Speech.

Tamariz (2008) Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3(2).

Tamariz, M. (2011). Could arbitrary imitation and pattern completion have bootstrapped human linguistic communication? *Interaction Studies,* 12(1).

Tamariz, M. and Smith, A.D.M. Smith (2008). Regularity in mappings between signals and meanings. In A.D.M. Smith, K. Smith and R. Ferrer i Cancho (eds.) *The Evolution of Language (EVOLANG 7)*. World Scientific.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Reading, Mass: Addison-Wesley.